

(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION  
EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété  
Intellectuelle  
Bureau international



(43) Date de la publication internationale  
25 mars 2004 (25.03.2004)

PCT

(10) Numéro de publication internationale  
**WO 2004/024654 A2**

(51) Classification internationale des brevets<sup>7</sup> : C07C

(74) Mandataires : BREESE, Pierre etc.; Breesé-Majerowicz,  
3, avenue de l'Opéra, F-75001 Paris (FR).

(21) Numéro de la demande internationale :  
PCT/FR2003/002676

(81) États désignés (national) : AE, AG, AL, AM, AT, AU, AZ,  
BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ,  
DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH,  
GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC,  
LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW,  
MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC,  
SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA,  
UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(22) Date de dépôt International :  
9 septembre 2003 (09.09.2003)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :  
02/11195 10 septembre 2002 (10.09.2002) FR

(84) États désignés (régional) : brevet ARIPO (GH, GM, KE,  
LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), brevet  
eurasien (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), brevet  
européen (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,  
FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK,  
TR), brevet OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,  
GW, ML, MR, NE, SN, TD, TG).

(71) Déposants (pour tous les États désignés sauf US) :  
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE - CNRS [FR/FR]; 3, rue Michel-Ange, F-75794  
Paris Cedex 16 (FR). UNIVERSITE DES SCIENCES  
ET TECHNOLOGIES DE LILLE (LILLE I) [FR/FR];  
Bâtiment M3, F-59655 Villeneuve D'ASQ Cedex (FR).

Publiée :

— sans rapport de recherche internationale, sera republiée  
dès réception de ce rapport

(72) Inventeur; et

(75) Inventeur/Déposant (pour US seulement) : VAN  
HOECKE, Marie-Pierre [FR/FR]; 4, rue Gambirinus,  
F-59520 Marquette-Lez-Lille (FR).

En ce qui concerne les codes à deux lettres et autres abréviations, se référer aux "Notes explicatives relatives aux codes et abréviations" figurant au début de chaque numéro ordinaire de la Gazette du PCT.

(54) Title: METHOD OF DETERMINING BRANCHED MOLECULES FROM MASS DATA

(54) Titre : PROCEDE DE DETERMINATION DE MOLECULES BRANCHEES A PARTIR DE DONNEES DE MASSE

(57) Abstract: The invention relates to a method of determining a branched molecular structure from data relating to masses of fragments of said molecule. The inventive method is characterised in that it comprises the following steps: (a) the list of basic elements that can form the branched molecule is saved in a memory element; (b) solutions to an equation are stored in a memory element, said equation involving the aforementioned basic elements, the mass of same, the number thereof and one of the given masses, step (b) being performed for all of the masses; (c) sequences of basic elements are created from said solutions, each sequence comprising a solution for a minimum mass and the full sequence being the solution for a maximum mass; (d) the sequences are grouped together by composition; (e) possible tree structures are stored for a composition of basic elements according to the sequences for said composition which were determined in step (c); (f) all of the possible fragments of the tree structure are calculated for each tree structure from step (e); and (g) a test is performed for each fragment from step (f) in order to ascertain whether or not said fragment corresponds to one of the given masses.

(57) Abrégé : Procédé de détermination d'une structure moléculaire branchée à partir de données de masses de fragments de ladite molécule, caractérisé en ce qu'il comprend les étapes suivantes a) une étape d'enregistrement dans une mémoire de la liste des éléments de base pouvant constituer ladite molécule branchée b) une étape de stockage en mémoire des solutions à une équation mettant en jeu les éléments de base, leur masse, leur nombre et une des masses données, ceci pour toutes les masses, c) une étape de constitution de séquences d'éléments de base à partir desdites solutions, chaque séquence incluant une solution pour une masse dite minimale et la séquence complète étant solution pour une masse dite maximale; d) une étape de regroupement des séquences par composition; e) une étape de stockage des arbres possibles pour une composition d'éléments de base en fonction des séquences de cette composition déterminées à l'étape c); f) pour chaque arbre de l'étape e), une étape de calcul de l'ensemble des fragments possibles de l'arbre; g) pour chaque fragment de l'étape f), une étape de test permettant de savoir si le fragment correspond à une des masses données.

WO 2004/024654 A2

PROCÉDÉ DE DÉTERMINATION DE MOLÉCULES BRANCHÉES À PARTIR DE  
DONNÉES DE MASSE

5           La présente invention se rapporte au domaine de  
l'étude de molécules et de détermination de leur  
composition et de leur structure. En particulier, la  
présente invention se rapporte à la détermination  
automatique de structures moléculaires branchées en  
10           utilisant des données de masse. Une application de la  
présente invention est la détermination de la structure  
d'oligosaccharides à partir de données de masse fournies  
par un spectromètre de masse.

15           Dans ce domaine, la technique habituellement  
utilisée est une étude manuelle des données fournies par le  
spectromètre de masse confrontées à une expertise humaine.  
Cette étude est très coûteuse en temps.

20           Des solutions ont donc été proposées pour  
réaliser de manière automatique l'étude des données de  
masse, mais les outils développés ne permettent pour le  
moment que de déterminer les structures linéaires.

25           Le problème technique que la présente invention  
entend résoudre est la détermination d'une structure  
moléculaire branchée à partir d'un spectre de masse ou  
d'autres données de masse, ceci de manière entièrement  
automatique sans intervention de l'homme. Les résultats de  
la détermination étant destinés à des experts, ceux-ci  
pourront infirmer ou confirmer les résultats donnés  
automatiquement.

30           La présente invention propose donc de déterminer  
automatiquement la structure branchée la plus probable pour  
une molécule, les structures linéaires étant incluses dans  
l'ensemble des structures branchées. Pour cela, la présente  
invention réalise un certain nombre d'opérations sur  
35           l'ensemble des masses fourni et délivre un résultat.

L'expertise humaine peut être requise pour orienter le processus ou valider la solution proposée par le procédé mais cette intervention n'est que ponctuelle et brève. Ainsi, le temps d'intervention de l'expert est limité aux  
5 seules questions nécessitant réellement une compétence scientifique.

Pour ce faire, la présente invention est du type décrit ci-dessus et elle est remarquable dans son  
10 acceptation la plus large, en ce qu'elle concerne un procédé de détermination d'une structure moléculaire branchée à partir de données de masses de fragments de ladite molécule, comprenant les étapes suivantes :

- 15 a) une étape d'enregistrement dans une mémoire de la liste des éléments de base pouvant constituer ladite molécule branchée ;
- b) une étape de stockage en mémoire des solutions à une équation mettant en jeu les éléments de base, leur masse, leur nombre et  
20 une des masses données, ceci pour toutes les masses ;
- c) une étape de constitution de séquences d'éléments de base à partir desdites solutions, chaque séquence incluant une  
25 solution pour une masse dite minimale et la séquence complète étant solution pour une masse dite maximale ;
- d) une étape de regroupement des séquences par composition ;
- 30 e) une étape de stockage des arbres possibles pour une composition d'éléments de base en fonction des séquences de cette composition déterminées à l'étape c) ;

f) pour chaque arbre de l'étape e), une étape de calcul de l'ensemble des fragments possibles de l'arbre ;

5 g) pour chaque fragment de l'étape f), une étape de test permettant de savoir si le fragment correspond à une des masses données ;

Avantageusement, l'étape b) est réalisée de manière incrémentale depuis la plus petite masse vers la plus grande masse, la solution pour une masse est cherchée  
10 en utilisant les solutions trouvées pour les masses inférieures et les données correspondant aux dites solutions sont stockées dans un tableau.

De préférence, l'étape c) consiste à définir le N-ème élément de base de la séquence en comparant la  
15 solution N pour la masse avec la solution N-1 à partir de laquelle la solution N a été trouvée et à écrire dans un fichier un identifiant dudit N-ème élément de base.

Selon un mode de réalisation préféré, l'étape e) consiste à :

- 20 - associer à chaque élément de base d'une séquence une donnée de type « nœud » comportant un identifiant de l'élément de base et au moins une référence à un autre nœud ;
- 25 - à la N-ème étape, pour chaque arbre de l'étape N-1, pour chaque nœud comportant une référence libre, créer un nœud contenant le composant N de la séquence et affecter ladite référence libre audit nœud créé.

30

Avantageusement, l'étape f) consiste à générer une liste de séquences d'éléments de base où chaque séquence inclut ladite solution pour une masse minimale, les éléments du fragment correspondant à ladite séquence

étant ordonnés par ajout de « nœud » en « nœud » à partir de ladite solution pour une masse minimale.

De préférence, l'étape g) est composée de deux étapes :

- 5           - une étape de comparaison de la séquence correspondant audit fragment avec les séquences de ladite composition résultant un premier booléen VRAI ou FAUX ;
- 10          - Si ledit premier booléen est FAUX, une étape de comparaison de la composition de ladite séquence avec les compositions des sous-séquences de même longueur incluant la solution minimale desdites séquences solutions pour une masse maximale résultant un deuxième booléen VRAI ou FAUX.

15           Selon un autre mode de réalisation, le procédé comprend une étape supplémentaire de choix de l'arbre (des arbres) le(s) plus pertinent(s) en fonction des résultats de l'étape g) en associant à chacun des arbres générés à l'étape e) un compteur mis à zéro au début du procédé et  
20           incrémenté d'un si lesdits deux booléens sont FAUX et en choisissant l'arbre (ou les arbres) dont le(s) compteur(s) est (sont) le(s) plus faible(s).

La présente invention se rapporte également à une utilisation du procédé de détermination d'une structure  
25           moléculaire branchée décrit dans les paragraphes précédents caractérisée en ce que la structure recherchée est un oligosaccharide, les données de masse sont obtenues par spectrométrie de masse et les éléments de base sont des monosaccharides ou des groupements substituants.

30           On comprendra mieux la présente invention à l'aide de la description, faite ci-après à titre purement explicatif, d'un mode de réalisation de l'invention, en référence aux figures annexées :

- 35           - La figure 1 représente un spectre arbitraire de masse simulant un spectre expérimental.

- La figure 2 illustre la première partie du déroulement d'un mode de réalisation de l'invention.
- La figure 3 illustre la deuxième partie du déroulement d'un mode de réalisation de l'invention.

Le procédé selon l'invention comporte 5 étapes précédées d'une étape préliminaire réalisée indépendamment du procédé :

L'étape préliminaire consiste à obtenir un ensemble de masses correspondant à des fragments de la molécule à déterminer. Cet ensemble de masses est appelé « spectre expérimental ».

La première étape consiste à enregistrer l'ensemble des molécules simples susceptibles de composer la molécule à déterminer.

La deuxième étape consiste à déterminer l'ensemble des chemins allant d'une structure racine à une structure finale où la structure racine correspond à une valeur dite « minimale » de l'ensemble des masses et la structure finale correspond à une valeur dite « maximale » de l'ensemble des masses. L'ensemble de ces chemins passe par des structures intermédiaires, c'est-à-dire incluant la structure racine et incluses dans la structure finale, et correspondant à des valeurs de masses comprises entre la valeur minimale et la valeur maximale.

La troisième étape consiste à générer des séquences de molécules simples obtenues à partir desdits chemins et de regrouper les séquences ayant les mêmes nombres de chaque molécule simple en « composition ».

La quatrième étape consiste à déterminer pour chaque composition, l'ensemble des arbres possibles. De préférence, chaque arbre doit pouvoir être constitué à partir de n'importe quelle séquence de la décomposition.

La cinquième étape consiste à calculer pour chaque arbre le « spectre » théorique de l'arbre en déterminant de tous les fragments possibles de l'arbre contenant la racine et à comparer le spectre théorique avec le spectre expérimental.

Le résultat de la comparaison permet de déterminer quel est l'arbre le plus probable.

Le procédé selon l'invention peut être utilisé pour déterminer de manière automatique la composition d'oligosaccharides. Pour la détermination d'un oligosaccharide, il comporte plusieurs étapes :

Une étape préliminaire est destinée à obtenir un ensemble de masses (appelé « spectre ») obtenu par spectrométrie de masse de la molécule à déterminer. Les masses de cet ensemble comprennent les masses de fragments de la molécule, de produits de recombinaison entre les composants de la molécule ou de fragments substitués.

Une première étape consiste à enregistrer dans une mémoire la liste des monosaccharides connus ainsi que leur masse.

Une deuxième étape consiste à parcourir l'ensemble des masses déterminées par spectrométrie. Pour une première masse, le procédé cherche à résoudre l'équation suivante, appelée équation Y :

masse totale mesurée =  
 (Somme des masses des composants - pertes de liaisons)  
 + (aglycone - perte liaison aglycone-root)\* + masse ION +  
 5 agent réducteur  
 (\*) ssi aglycone ≠ 0

Cette équation se traduit par =  

$$M \pm err = \sum (ai * mi) - \left[ \left( \sum (ai) - 1 \right) * H_2O \right] + (aglycone - H_2O)^* + ION + reduction$$
  
 (\*): ssi aglycone ≠ 0

10 où :

M = masse expérimentale mesurée dans le spectromètre de masse

err = erreur de mesure du spectromètre

mi = masse (monoisotopique) du composant i

15 ai = nombre de composants i apparaissant dans la solution  
 (ai est un entier)

H<sub>2</sub>O = masse d'une molécule d'EAU

aglycone = masse de l'aglycone en cas d'aminoréduction

ION = masse de l'ion

20 reduction = incrément de masse dû aux conditions de réduction

Selon le procédé, l'ensemble des masses déterminées est parcouru dans l'ordre des masses  
 25 croissantes et pour chaque masse, on cherche un ou plusieurs monosaccharides résolvant l'équation Y. La plus petite masse pour laquelle l'équation Y a une solution est appelée masse minimale et la solution à l'équation Y pour la masse minimale est appelée « racine ». La racine peut  
 30 être composée d'un ou de plusieurs monosaccharides. Cette racine est le premier élément d'un ensemble de chemins : l'ensemble des chemins est un tableau et chaque ligne du tableau constitue une étape d'un chemin.



Par la suite, on continue le parcours des masses par ordre croissant en essayant de résoudre l'équation Y avec des structures de monosaccharides incluant la racine. Chaque structure solution est ajoutée audit tableau. De manière étendue, chaque structure solution de l'équation Y pour une des masses mesurées inclut une structure préalablement enregistrée dans le tableau. Ainsi, un système d'héritage est mis en place à partir de la racine : chaque structure solution, exceptée la racine, a une « mère » parmi les autres structures solutions.

Dans ce tableau, chaque ligne correspond à une structure de monosaccharide et donne la quantité de chaque monosaccharide (supérieur ou égal à 0) dans la structure ainsi que le numéro de la ligne de la structure « mère » de la structure courante.

La recherche d'une solution de l'équation Y pour une masse dite courante consiste donc à ajouter au moins un monosaccharide à une structure solution de l'équation Y pour une masse inférieure à la masse courante. Pour cela, ledit tableau est parcouru et pour chaque ligne (i.e. chaque structure), une solution incluant la structure correspondante à ladite ligne est cherchée. Afin de réduire le temps de calcul, certaines lignes ne sont pas traitées : les structures solutions de l'équation Y pour une masse inférieure d'une certaine quantité à la masse courante ne sont pas incluses dans la recherche. Cette quantité est choisie arbitrairement par l'utilisateur. Dans un mode de réalisation du procédé, cette quantité était égale à deux fois la masse du monosaccharide le plus lourd (NeuGC).

Ainsi, soit une structure solution S1 de l'équation Y pour une masse de raies r1, s'il n'existe aucun monosaccharide ou assemblage de monosaccharides, qui, agrégé à ladite solution S1 est solution de l'équation Y pour toutes les masses expérimentales comprises entre r1 et  $r2=r1+2*Masse(NeuGC)$ , alors la solution S1 n'a pas de fils

et n'est plus prise en compte pour les recherches des solutions de l'équation Y pour des masses supérieures à r2.

Dans l'exemple de la figure 1, la première raie correspond à une masse de 300,4 Daltons.

5            r1 = 300,4 Da  
            M(NeuGC) = 327,1165 Da  
            2\*M(NeuGC) = 654,233 Da  
            r1 + 2\*M(NeuGC) = 954,633 Da

10           r2 = 500,30 Da < 954,633 Da  
            r3 = 665,50 Da < 954,633 Da  
            r4 = 811,56 Da < 954,633 Da  
            r5 = 827,80 Da < 954,633 Da  
            r6 = 973,86 Da > 954,633 Da

15

Ainsi pour les raies 2 à 5, on cherchera à combiner la solution pour la raie 1 avec une ou plusieurs molécules de base. En revanche, pour la raie 6, on ne cherchera pas à combiner la solution pour la raie 1 avec  
20           une ou plusieurs molécules de base.

La masse maximale pour laquelle Y a une solution est appelée « masse maximale ». Seuls les chemins aboutissant à une structure solution de l'équation Y pour  
25           la masse maximale sont considérés comme valables.

Une troisième étape intervient une fois l'ensemble des structures solutions déterminées : les structures solutions de l'équation Y pour la masse maximale  
30           sont traitées. En effet, seules ces structures sont susceptibles de correspondre à la molécule recherchée car elles couvrent tout le spectre de la raie minimale (racine) à la raie maximale pour laquelle elles sont solutions. L'étude de l'héritage des structures sélectionnées permet  
35           d'identifier la séquence des monosaccharides, c'est-à-dire

l'ordre dans lequel ils ont été ajoutés à la racine. On obtient ainsi un ensemble de séquences que l'on stocke dans une mémoire.

5 Certaines de ces séquences ont la même composition, c'est-à-dire la même quantité de chaque monosaccharide ou de chaque groupement substituant. Ces séquences de même composition sont regroupées en une seule « composition » de l'équation Y pour la masse maximale.

10 Pour chaque « composition », une quatrième étape consiste à déterminer les arbres possibles. Pour cela, l'utilisateur détermine pour une première séquence de ladite composition les arbres possibles :

- 15 - chaque élément de la séquence (un monosaccharide) est associé à un « nœud » qui comprend trois liens vers trois autres « nœuds » et un identifiant de l'élément. Ces liens sont appelés gauche, droite et milieu ;
- 20 - ainsi, le premier élément de la séquence (la racine) est associé à un premier nœud ;
- 25 - pour le deuxième élément de la séquence, on crée trois ensembles de nœuds : chaque ensemble contient deux nœuds dont le premier correspond à la racine et le second audit deuxième élément, les deux nœuds étant respectivement liés par le lien gauche, le lien droite et le lien milieu pour les premier, deuxième et troisième ensembles. Ces ensembles de nœuds sont appelés des « arbres » et l'ensemble (2) des arbres
- 30 contient les arbres comprenant le deuxième élément ;
- Ainsi de suite, pour le n-ième élément, l'ensemble (n) des arbres est composé d'arbres créés à partir des arbres de l'ensemble (n-1),
- 35 chaque nouvel arbre correspondant à un arbre de

l'ensemble (n-1) ajouté d'un nœud correspondant au n-ième élément sur un des liens libres.

5 - Pour réduire le temps de calcul de l'ensemble des arbres final, on supprime au fur et à mesure les arbres redondants : par exemple, les trois arbres composés de deux molécules de fucose où la deuxième molécule est située respectivement sur les liens droite, gauche et milieu, sont équivalents. Ainsi, un certain nombre d'arbres  
10 sont éliminés.

Ensuite, les arbres restants sont comparés aux autres séquences de la même composition : un arbre est conservé si toutes les séquences de la même composition peuvent être réalisées avec cet arbre.

15 Le choix de trois liaisons possibles à partir d'un nœud a été pris en référence à la valence 4 de l'atome de carbone sur lequel se fixe en général l'élément de base suivant.

20 Pour une composition, il reste donc un ensemble d'arbres « compatibles » avec toutes les séquences de la composition.

25 Afin de déterminer quel est l'arbre le plus probable de manière automatique, le procédé propose dans une cinquième étape de comparer le spectre théorique de chaque arbre restant avec le spectre expérimental mesuré par le spectromètre de masse. Pour cela, le procédé compte  
30 le nombre de raies du spectre théorique qui n'ont pas pu être utilisées par le procédé. Une raie du spectre théorique d'un arbre correspond à la masse d'un fragment de l'arbre. Le calcul du spectre théorique d'un arbre revient donc à calculer les masses des sous-arbres inclus dans  
35 l'arbre et contenant la racine. Le nombre de masses de

sous-arbres n'existant pas dans l'ensemble des masses expérimentales détermine la probabilité d'occurrence de l'arbre en question.

5 La méthode employée de préférence par le procédé permet de réduire le temps de calcul : pour un arbre, le calcul de la liste des fragments se fait de la manière suivante :

- 10 - un opérateur « multiplication d'une liste par un élément » est créé qui, à partir d'une liste d'éléments, crée une nouvelle liste où chaque élément est le résultat de la concaténation de l'élément nouveau avec un élément de la liste d'entrée.
- 15 - un opérateur « produit de deux listes » résulte du premier : c'est l'application de l'opérateur « multiplication d'une liste par un élément » sur tous les items de la liste 1 avec la liste 2.

20 Ainsi, la liste des fragments produite à un nœud quelconque est égale au produit des listes issues de ses fils, qui est ensuite multiplié par l'élément du nœud ; une masse nulle est ajoutée enfin en tête de liste ; l'introduction de la masse nulle implémente le fait que la  
25 branche peut-être absente ; cette masse nulle se propage dans le parcours récursif et permet d'avoir la liste complète en un seul parcours ; la liste produite par une feuille est donc une liste de deux éléments : [0, elt].

30 Une fois la liste des fragments de l'arbre obtenue, le procédé détermine le nombre de fragments théoriques trouvés ne correspondant à aucune des masses expérimentales fournies. Afin d'éviter de recalculer la masse théorique pour chaque fragment, le procédé propose de comparer les fragments théoriques déterminés avec les  
35 décompositions d'une « séquence ». En effet, la liste de

fragments construite est composée de fragments présentés sous forme de suite de monosaccharides. Pour un fragment, si ladite suite de monosaccharides est présente dans une des décompositions de la « séquence », alors il existe une

5 raie du spectre correspondant à cette suite de monosaccharides. Donc le fragment théorique correspondant est présent dans le spectre expérimental. De plus, il peut arriver que la suite de monosaccharides représentant un

10 fragment ne soit pas ordonnée de façon à ce qu'elle soit reconnue comme valable. Pour résoudre ce genre de cas, le procédé réalise une comparaison des compositions de la suite de monosaccharides, sans ordre, avec la partie de même taille des décompositions. Si les deux compositions sont identiques, le fragment correspond à une raie du

15 spectre expérimental.

Les fragments dont la composition ne se retrouve pas parmi les décompositions sont appelées des « raies manquantes » du spectre théorique de l'arbre. Le nombre de raies manquantes détermine la pertinence de l'arbre.

20 L'arbre ayant le moins de raies manquantes définit la structure la plus probable pour la molécule. Si plusieurs arbres ont le même nombre minimal de raies manquantes, il est nécessaire de recourir à une expertise humaine qui saura déterminer quel est l'arbre le plus probable.

25 En particulier, cette expertise s'appuie sur l'équilibre naturel des molécules. Une extension du procédé de l'invention peut prendre en compte cet équilibre pour déterminer l'arbre le plus probable, en comptant par exemple le nombre de monosaccharides sur chaque sous-arbre

30 d'un nœud comportant plusieurs sous-arbres ainsi que le type des monosaccharides.

Un exemple de réalisation de ce procédé est décrit ci-dessous en se référant aux dessins.

Le spectromètre de masse fournit les données représentées sur la figure 1, où chaque pic (ou raie) correspond à la masse d'un fragment de la molécule. On considère que l'oligosaccharide cherché est composé de  
 5 HexNAC (masse : 221,0899 Da), d'Hexose (masse : 180,0364 Da) et de Fucose (masse : 164,0684 Da). La résolution de l'équation Y donne le tableau suivant :

N° ligne	HexNAC	Hexose	Fucose	Ligne « mère »	N° raie
1	1	0	0	-	1
2	2	0	0	1	2
3	2	0	1	2	3
4	2	1	0	2	4
5	2	1	1	3	5
6	2	1	1	4	5

10 Il y a deux solutions pour la raie maximale : en remontant le chemin menant de la racine (raie 1) à la raie maximale, on obtient deux séquences :

HexNAC-HexNAC-Fucose-Hexose (séquence 1)

HexNAC-HexNAC-Hexose-Fucose (séquence 2)

15 Ces deux séquences ou décompositions ont la même composition, elles sont donc regroupées dans une seule solution.

On cherche maintenant les arbres possibles pour la séquence 2. La construction des arbres est illustrée  
 20 figure 2. Une première étape consiste à créer un arbre contenant un premier « HexNAC ». la deuxième étape consiste à ajouter un deuxième HexNAC audit premier HexNAC. Le deuxième HexNAC peut être accroché au premier par le lien « gauche », le lien « milieu » ou le lien « droit ». Dans  
 25 la pratique, comme ces trois arbres sont équivalents, un seul arbre est construit, avec le deuxième hexNAC accroché sur le lien « gauche ». D'une manière générale, un nouveau

monosaccharide sera toujours accroché sur le lien libre le plus à gauche du nœud précédent et un seul arbre sera construit quel que soit le nombre de liens libres du nœud. La troisième étape consiste à ajouter un Hexose à l'arbre construit à l'étape 2 : pour cela, il y a deux possibilités non équivalentes :

- accrocher l'Hexose au premier hexNAC ;
- accrocher l'Hexose au deuxième hexNAC ;

Ainsi deux arbres sont construits.

10 La quatrième étape consiste enfin à ajouter le Fucose aux arbres construits à l'étape 3. Les six possibilités (trois par arbre) sont détaillées sur la figure 2. Il est à noter que la molécule HexNAC qui se situait sur le lien gauche du premier HexNAC du premier arbre de l'étape 3 est maintenant  
15 située sur le lien milieu pour l'arbre N°5 de l'étape 4. En effet, sur un nœud, les sous-arbres sont triés de gauche à droite par ordre de poids décroissant. Comme l'association d'un Hexose et d'un Fucose est plus lourde qu'HexNAC, l'ordre est inversé par rapport aux autres arbres  
20 possibles, où ce cas ne se présente pas.

Une fois les arbres construits pour la séquence 2, le procédé selon l'invention vérifie que les arbres construits sont compatibles avec la séquence 1. Pour cela, le procédé teste s'il est possible de reconstruire les  
25 arbres de l'étape 4 avec la séquence 1. Deux arbres sont éliminés (les arbres N°5 et 6) car il est impossible de construire ces arbres sans placer l'Hexose avant le Fucose.

Sur les 4 arbres restants, le procédé selon l'invention cherche à déterminer le spectre théorique afin  
30 de le comparer avec le spectre expérimental. Les fragments sont déterminés selon la méthode décrite ci-dessus utilisant les opérateurs « multiplication d'une liste par un élément » et « produit de deux listes ». Chaque fragment déterminé est décrit sous forme d'une séquence stockée dans  
35 une mémoire. Cette méthode est illustrée figure 3.



Par exemple, pour le premier arbre, le procédé crée trois listes, chaque liste correspondant à un des « fils » du nœud racine :

- 5           - (HexNAC,  $\emptyset$ ) ;  
          - (Hexose,  $\emptyset$ ) ;  
          - (Fucose,  $\emptyset$ ).

On applique l'opérateur « produit de deux listes » aux deux premières listes, ce qui donne :

10           (HexNAC-Hexose, HexNAC, Hexose,  $\emptyset$ )  
          que l'on multiplie par la troisième liste,  
          soit :

(HexNAC-Hexose-Fucose, HexNAC-Hexose, HexNAC-Fucose, HexNAC, Hexose-Fucose, Hexose, Fucose,  $\emptyset$ )

15

On applique l'opérateur « multiplication d'une liste par un élément » à la liste précédente avec l'élément « HexNAC », ce qui donne :

20           (HexNAC-HexNAC-Hexose-Fucose, HexNAC-HexNAC-Hexose, HexNAC-HexNAC-Fucose, HexNAC-HexNAC, HexNAC-Hexose-Fucose, HexNAC-Hexose, HexNAC-Fucose, HexNAC)

Cette liste est la liste des fragments pour le premier arbre. Chaque élément de cette liste correspond à une raie du spectre théorique de l'arbre concerné. Pour  
25 vérifier que les raies théoriques existent dans le spectre expérimental, il suffit de vérifier que le fragment correspondant est inclus dans une des « séquences » de la « décomposition ». Ces séquences étaient :

30           HexNAC-HexNAC-Fucose-Hexose (séquence 1)  
          HexNAC-HexNAC-Hexose-Fucose (séquence 2)

Ainsi, en numérotant les fragments de la liste de 1 à 8, on constate que les fragments 1, 2, 3, 4 et 8 sont inclus dans une des séquences, alors que les fragments 5, 6 et 7 ne le sont pas. La deuxième vérification consiste  
35 à regarder la composition des fragments non-valables avec

la composition des fragments de même longueur, contenant la racine des « séquences ». Dans ce cas, les trois fragments, sont également rejetés. Ainsi, le nombre de raies manquantes de cet arbre est de 3. Les mêmes étapes sont

5 réalisées pour les autres arbres. L'arbre qui a le plus petit nombre de raies manquantes est la plus probable, en l'occurrence ici, le quatrième.

10 L'invention est décrite dans ce qui précède à titre d'exemple. Il est entendu que l'homme du métier est à même de réaliser différentes variantes de l'invention sans pour autant sortir du cadre du brevet

REVENDICATIONS

1. Procédé de détermination d'une structure  
5 moléculaire branchée à partir de données de masses de fragments de ladite molécule, caractérisé en ce qu'il comprend les étapes suivantes :

- 10 a) une étape d'enregistrement dans une mémoire de la liste des éléments de base pouvant constituer ladite molécule branchée ;
- b) une étape de stockage en mémoire des solutions à une équation mettant en jeu les éléments de base, leur masse, leur nombre et une des masses données, ceci pour toutes les  
15 masses ;
- c) une étape de constitution de séquences d'éléments de base à partir desdites solutions, chaque séquence incluant une solution pour une masse dite minimale et la  
20 séquence complète étant solution pour une masse dite maximale ;
- d) une étape de regroupement des séquences par composition ;
- e) une étape de stockage des arbres possibles  
25 pour une composition d'éléments de base en fonction des séquences de cette composition déterminées à l'étape c) ;
- f) pour chaque arbre de l'étape e), une étape de calcul de l'ensemble des fragments possibles  
30 de l'arbre ;
- g) pour chaque fragment de l'étape f), une étape de test permettant de savoir si le fragment correspond à une des masses données.

2. Procédé de détermination d'une structure moléculaire branchée selon la revendication 1, caractérisé en ce que l'étape b) est réalisée de manière incrémentale depuis la plus petite masse vers la plus grande masse, que  
5 la solution pour une masse est cherchée en utilisant les solutions trouvées pour les masses inférieures et que les données correspondant aux dites solutions sont stockées dans un tableau.

3. Procédé de détermination d'une structure  
10 moléculaire branchée selon l'une des revendications précédentes, caractérisé en ce que l'étape c) consiste à définir le N-ème élément de base de la séquence en comparant la solution N pour la masse en cours de traitement avec la solution N-1 à partir de laquelle la solution N a été  
15 trouvée et à écrire dans un fichier un identifiant dudit N-ème élément de base.

4. Procédé de détermination d'une structure moléculaire branchée selon l'une des revendications précédentes, caractérisé en ce que l'étape e) consiste à :

- 20 - associer à chaque élément de base d'une séquence une donnée de type « nœud » comportant un identifiant de l'élément de base et au moins une référence à un autre nœud ;
- 25 - à la N-ème étape, pour chaque arbre de l'étape N-1, pour chaque nœud comportant une référence libre, créer un nœud contenant le composant N de la séquence et affecter ladite référence libre audit nœud créé ;

30 5. Procédé de détermination d'une structure moléculaire branchée selon l'une des revendications précédentes, caractérisé en ce que l'étape f) consiste à générer une liste de séquences d'éléments de base où chaque séquence inclut ladite solution pour une masse minimale, les  
35 éléments du fragment correspondant à ladite séquence étant

ordonnés par ajout de « nœud » en « nœud » à partir de ladite solution pour une masse minimale.

6. Procédé de détermination d'une structure moléculaire branchée selon l'une des revendications précédentes, caractérisé en ce que l'étape g) est composé de deux étapes :

- Une étape de comparaison de la séquence correspondant audit fragment avec les séquences de ladite composition résultant un premier booléen VRAI ou FAUX ;
- Si ledit premier booléen est FAUX, une étape de comparaison de la composition de ladite séquence avec les compositions des sous-séquences de même longueur incluant la solution minimale desdites séquences solutions pour une masse maximale résultant un deuxième booléen VRAI ou FAUX.

7. Procédé de détermination d'une structure moléculaire branchée selon la revendication 6, caractérisé en ce qu'il comprend une étape supplémentaire de choix de l'arbre (des arbres) le(s) plus pertinent(s) en fonction des résultats de l'étape g) en associant à chacun des arbres générés à l'étape e) un compteur mis à zéro au début du procédé et incrémenté d'un si lesdits deux booléens sont FAUX et en choisissant l'arbre (ou les arbres) dont le(s) compteur(s) est (sont) le(s) plus faible(s).

8. Utilisation du procédé de détermination d'un structure moléculaire branchée selon l'une des revendications précédentes, caractérisée en ce que la structure recherchée est un oligosaccharide, les données de masse sont obtenues par spectrométrie de masse et les éléments de base sont des monosaccharides ou des groupements substituants.

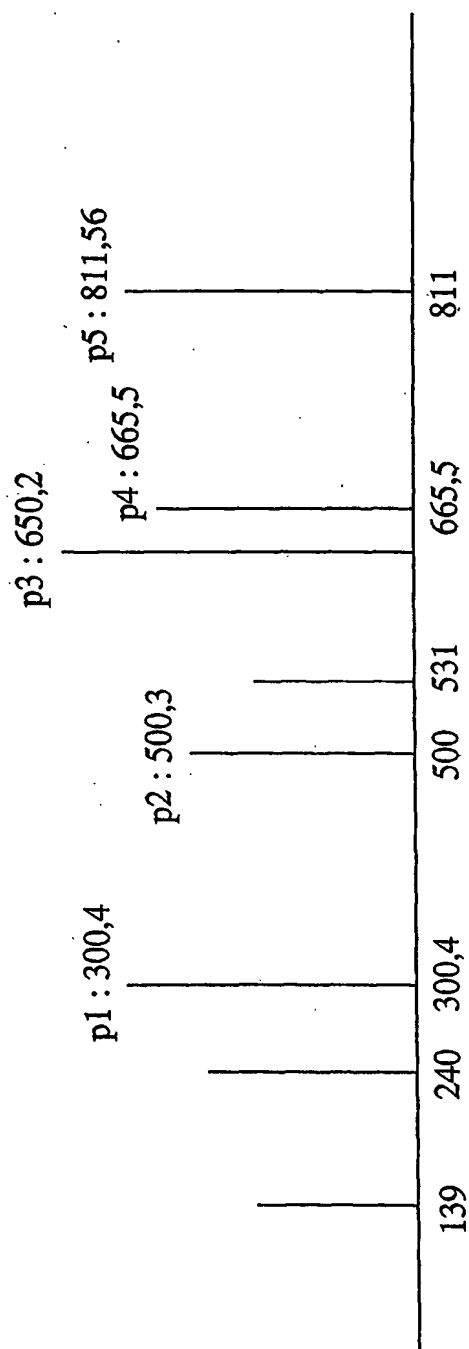


Figure 1

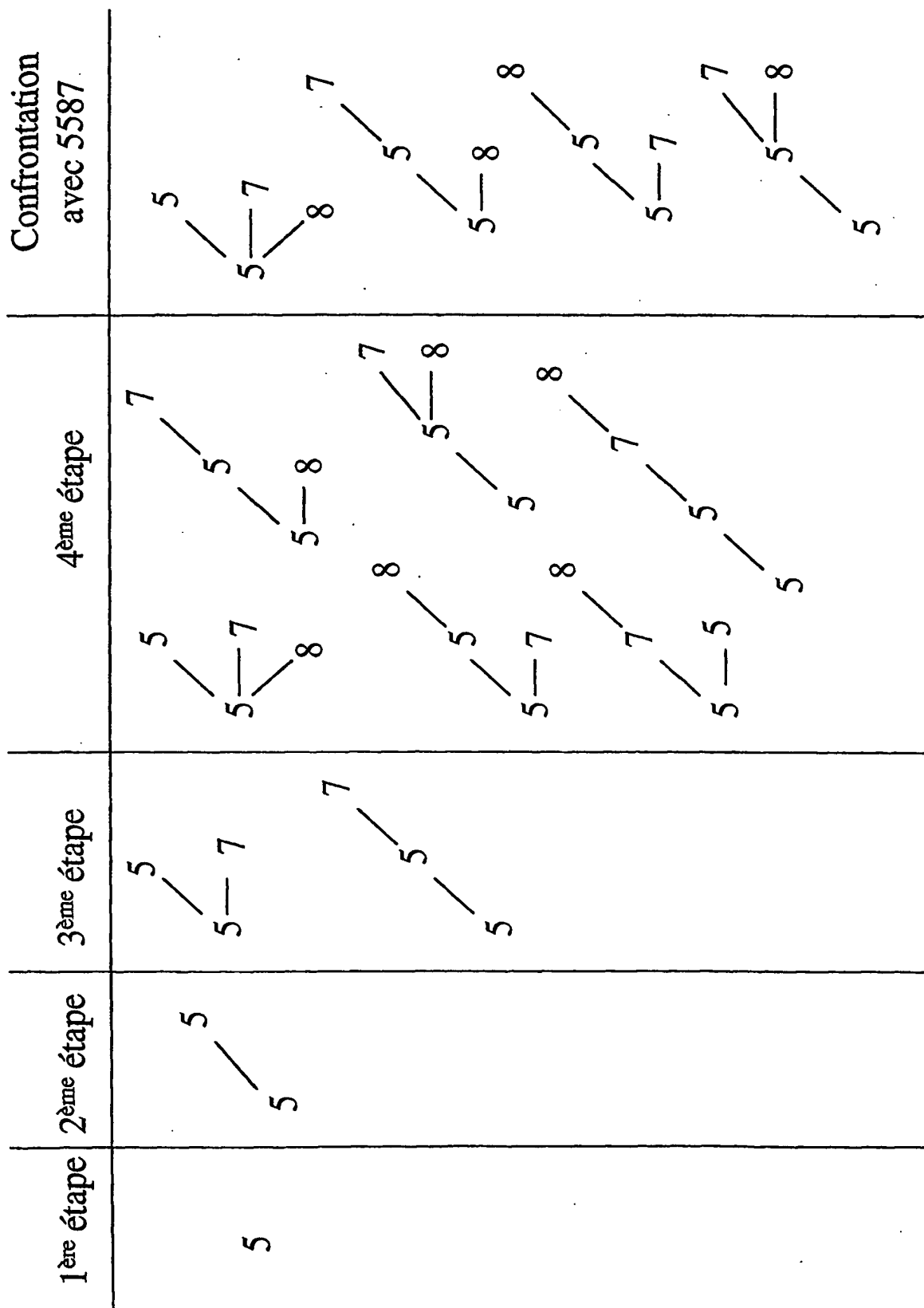


Figure 2

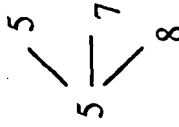
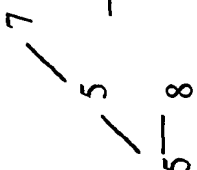
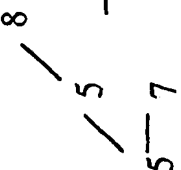

Arbres possibles	Fragments théoriques	Validité des fragments par rapport aux décompositions	Validité des fragments sans ordre	Nombre de raies manquantes
	5578, 557, 558, 55, 578, 57, 58, 5	✓ 5578, 557, 558, 55, 5 ✗ 578, 57, 58	✓ ∅ ✗ 578, 57, 58	3
	5587, 558, 557, 55, 57, 5	✓ 5587, 558, 557, 5 ✗ 57	✓ ∅ ✗ 57	1
	5578, 557, 558, 55, 58, 5	✓ 5578, 557, 558, 55, 5 ✗ 58	✓ ∅ ✗ 58	1
	5578, 557, 558, 55, 5	✓ 5578, 557, 558, 55, 5 ✗ ∅		0

Figure 3